

Dynamic trees for streaming and massive data contexts

Christoforos Anagnostopoulos*

Department of Mathematics, Imperial College London
South Kensington Campus, SW7 2AZ.
canagnos@imperial.ac.uk

Robert B. Gramacy

Booth School of Business, The University of Chicago
5807 S. Woodlawn Ave., Chicago, IL, 60637.
rbgramacy@ChicagoBooth.edu

January 27, 2012

Abstract

Data collection at a massive scale is becoming ubiquitous in a wide variety of settings, from vast offline databases to streaming real-time information. Learning algorithms deployed in such contexts must rely on single-pass inference, where the data history is never revisited. In streaming contexts, learning must also be temporally adaptive to remain up-to-date against unforeseen changes in the data generating mechanism. Although rapidly growing, the online Bayesian inference literature remains challenged by massive data and transient, evolving data streams. Non-parametric modelling techniques can prove particularly ill-suited, as the complexity of the model is allowed to increase with the sample size. In this work, we take steps to overcome these challenges by porting standard streaming techniques, like data discarding and downweighting, into a fully Bayesian framework via the use of informative priors and active learning heuristics. We showcase our methods by augmenting a modern non-parametric modelling framework, dynamic trees, and illustrate its performance on a number of practical examples. The end product is a powerful streaming regression and classification tool, whose performance compares favourably to the state-of-the-art.

1 Introduction

In online inference, the objective is to develop a set of update equations that incorporate novel information as it becomes available, without needing to revisit the data history. This results in model fitting algorithms whose space and time complexity remains constant as information accumulates, and can hence operate in streaming environments featuring continual data arrival, or navigate massive datasets sequentially. Such operational constraints are becoming imperative in certain application areas as the scale and real-time nature of modern data collection continues to grow.

*Corresponding author.

In certain simple cases, online estimation without information loss is possible via exact recursive update formulae, e.g., via conjugate Bayesian updating (see Section 3.1). In parametric dynamic modelling, approximate samples from the filtering distribution for a variable of interest may be obtained online via sequential Monte Carlo (SMC) techniques, under quite general conditions. In Taddy et al. (2011), SMC is used in a non-parametric context, where a ‘particle cloud’ of *dynamic trees* are employed to track parsimonious regression and classification surfaces as data arrive sequentially. However, the resulting algorithm is not, strictly speaking, *online*, since tree moves may require access to the full data history, rather than parametric summaries thereof. This complication arises as an essential by-product of non-parametric modelling, wherein the complexity of the estimator is allowed to increase with the dataset size. Therefore, this article recognises that maintaining constant operational cost as new data arrives necessarily requires discarding some (e.g., historical) data.

Specifically, and to help set notation for the remainder of the paper, we consider supervised learning problems with labelled data (\mathbf{x}_t, y_t) , for $t = 1, 2, \dots, T$, where T is either very large or infinite. We consider responses y_t which are real-valued (i.e., regression) or categorical (classification). The p -dimensional predictors \mathbf{x}_t may include real-valued features, as well as binary encodings of categorical ones. The dynamic tree model, reviewed shortly in Section 2, allows sequential non-parametric learning via local adaptation when new data arrive. However its complexity, and thus computational time/space demands, may grow with the data size t . The only effective way to limit these demands is to sacrifice degrees-of-freedom (DoF) in representation of the fit, and the simplest way to do that is to discard data; that is, to require the trees to work with a subset $w \ll t$ of the data seen so far.

Our primary concern in this paper is managing the information loss entailed in data discarding. First, we propose datapoint *retirement* (Section 3), whereby discarded datapoints are partially ‘remembered’ through conjugate informative priors, updated sequentially. This technique is well-suited to trees, which combine non-parametric flexibility with simple parametric models and conjugate priors. Nevertheless, forming new partitions in the tree still requires access to actual datapoints, and consequently data discarding comes at a cost of both information and flexibility. We show that these costs can be managed, to a surprising extent, by employing the right retirement scheme even when discarding data randomly. In Section 4, we further show that borrowing active learning heuristics to prioritise points for retirement, i.e., *active discarding*, leads to better performance still.

An orthogonal concern in streaming data contexts is the need for temporal adaptivity when the concept being learned exhibits *drift*. This is where the data generating mechanism evolves over time in an unknown way. Recursive update formulae will generally require modification to acquire temporally adaptive properties. One common approach is the use of exponential downweighting, whereby the contribution of past datapoints to the algorithm is smoothly downweighted via the use of *forgetting factors*. In Section 5 we demonstrate how historical data retirement via suitably constructed informative priors can reproduce this effect in the non-parametric dynamic tree modelling context, while remaining fully online. Using synthetic as well as real datasets, we show how this approach compares favourably against modern alternatives. The paper concludes with a discussion in Section 6.

2 Dynamic Trees

Dynamic trees (DTs) (Taddy et al., 2011) are a process-analog of Bayesian treed models (Chipman et al., 1998, 2002). The model specification is amenable to fast sequential inference by SMC, yielding a predictive surface which organically increases in complexity as more data arrive. Software is available in the `dynaTree` package (Gramacy and Taddy, 2011) for R on CRAN (R Development Core Team, 2010), which has been extended to cover the techniques described in this paper. We now review model specification and inference in turn.

2.1 Bayesian static treed models

Trees partition the input space \mathcal{X} into hyper-rectangles, referred to as *leaves*, using nested logical rules of the form $(x_i \geq c)$. For instance, the partition $(x_1 \geq 3) \cap (x_2 < -1)$, $(x_1 > 3) \cap (x_2 \geq -1)$ and $(x_1 < 3)$ represent a tree with one internal node, $(x_2 \geq -1)$, and three leaves. We denote by $\eta(\mathbf{x})$ the unique leaf where \mathbf{x} belongs to, for any $\mathbf{x} \in \mathcal{X}$.

Treed models condition the likelihood function on a tree \mathcal{T} and fit one instance of a given simple parametric model per leaf. In this way, a flexible model is built out of simple parametric models $(\theta_\eta)_{\eta \in \mathcal{L}_\mathcal{T}}$, where $\mathcal{L}_\mathcal{T}$ is the set of leaves in \mathcal{T} . This flexibility comes at the price of a hard model search and selection problem: that of selecting a suitable tree structure. In the seminal work of Chipman et al. (1998), a Bayesian solution to this problem was proposed that relied on a generative prior distribution over trees: a leaf node η may split with probability $p_{\text{split}}(\mathcal{T}, \eta) = \alpha(1 + D_\eta)^{-\beta}$, where $\alpha, \beta > 0$, and D_η is the depth of η in the tree \mathcal{T} . This induces a joint prior via the probability that internal nodes $\mathcal{I}_\mathcal{T}$ have split and leaves $\mathcal{L}_\mathcal{T}$ have not: $\pi(\mathcal{T}) \propto \prod_{\eta \in \mathcal{I}_\mathcal{T}} p_{\text{split}}(\mathcal{T}, \eta) \prod_{\eta \in \mathcal{L}_\mathcal{T}} [1 - p_{\text{split}}(\mathcal{T}, \eta)]$. The specification is completed by employing independent sampling models at the tree leaves: $p(y_1, \dots, y_n | \mathcal{T}, \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{\eta \in \mathcal{L}_\mathcal{T}} p(y^\eta | \mathcal{T}, \mathbf{x}^\eta)$. Sampling from the posterior proceeds by MCMC, via proposed local changes to \mathcal{T} : so-called *grow*, *prune*, *change*, and *swap* “moves”. Any data type/model may be used as long as the marginal likelihoods $p(y^\eta | \mathcal{T}, \mathbf{x}^\eta)$ are analytic, i.e., as long as their parameters can be integrated out. This is usually facilitated by fully conjugate, scale invariant, default (non-informative) priors, e.g.,:

$$y | \mathbf{x} \sim N\left(\beta_{\eta(\mathbf{x})}^T \mathbf{x} + \mu_{\eta(\mathbf{x})}, \sigma_{\eta(\mathbf{x})}^2\right), \quad \pi(\beta_{\eta(\mathbf{x})}, \mu_{\eta(\mathbf{x})}, \sigma_{\eta(\mathbf{x})}^2) \propto \frac{1}{\sigma_{\eta(\mathbf{x})}^2} \quad (1)$$

for linear, or, letting $\beta_\eta = \mathbf{0}$, constant regression leaves. Similarly, multinomial leaves for classification with Dirichlet priors can be employed. These choices yield analytical posteriors (Taddy et al., 2011) but also efficient recursive updates for incorporating new datapoints (see Section 3.1).

2.2 Dynamic Trees

In DTs the “moves” are embedded into a process, which describes how old trees mature into new ones when new data arrive. Suppose that \mathcal{T}_{t-1} represents a set of recursive partitioning rules associated with \mathbf{x}^{t-1} , the set of covariates observed up-to time $t - 1$. The fundamental insight underlying the DT process is to view this tree as a *latent state*, evolving according to a state transition probability, $P(\mathcal{T}_t | \mathcal{T}_{t-1}, \mathbf{x}_t)$. The dependence on \mathbf{x}_t (but not on y_t) allows us to consider only moves *local to the current observation*: i.e., pruning or growing can only

occur (if at all) for the leaf $\eta(\mathbf{x}_t)$. This builds computational tractability into the process, as we eitherway need to recompute in that area. Formally, we let:

$$P(\mathcal{T}_t \mid \mathcal{T}_{t-1}, \mathbf{x}_t) = \begin{cases} 0, & \text{if } \mathcal{T}_t \text{ is not reachable from } \mathcal{T}_{t-1} \text{ via moves local to } \mathbf{x}_t \\ p_m \pi(\mathcal{T}_t), & \text{otherwise.} \end{cases} \quad (2)$$

where p_m is the probability of the unique move that can produce \mathcal{T}_t from \mathcal{T}_{t-1} , and π is the tree prior. We allow three types of moves: grow, prune and stay moves. Each type is considered equiprobable, whereas for grow moves, we choose among all possible split locations by first choosing a dimension j uniformly at random, and splitting $\eta(\mathbf{x}_t)$ around the location $x_j = \xi$ chosen uniformly at random from the interval formed from the projection of $\eta(\mathbf{x}_t)$ on the j th input dimension. The new observation, y_t , completes a stochastic rule for the update $\mathcal{T}_{t-1} \rightarrow \mathcal{T}_t$ via $p(y^t \mid \mathcal{T}_t, \mathbf{x}^t)$ for each $\mathcal{T}_t \in \{\mathcal{T}_t\}$.

The DT specification is amenable to Sequential Monte Carlo (e.g., Carvalho et al., 2010) inferential mechanics. At each iteration t , the discrete approximation to the tree posterior $\{\mathcal{T}_{t-1}^{(i)}\}_{i=1}^N$, based on N particles, can be updated to $\{\mathcal{T}_t^{(i)}\}_{i=1}^N$ by *resampling* and then *propagating*. Resampling the particles (with replacement) proceeds according to their predictive probability for the next (\mathbf{x}, y) pair, $w_i = p(y_t \mid \mathcal{T}_{t-1}^{(i)}, \mathbf{x}_t)$. Then, propagating each resampled particle follows the process outlined in 2.2. Both steps are computationally efficient because they involve only local calculations (requiring only the subtrees of the parent of each $\eta^{(i)}(\mathbf{x})$). Nevertheless, the particle approximation can shift great distances in posterior space after an update because the data governed by $\eta(\mathbf{x}_t)^{(i)}$ may differ greatly from one particle to another, and thus so may the weights w_i . This appealing division of labour mimicks the behaviour of an ensemble method without explicitly maintaining one. As with all particle simulation methods, some Monte Carlo (MC) error will accumulate and, in practice, one must be careful to assess its effect. Nevertheless, DT out-of-sample performance compares favourably to other nonparametric methods, like Gaussian processes (GPs) regression and classification, but at a fraction of the computational cost (Taddy et al., 2011).

3 Datapoint retirement

At time t , the DT algorithm of Taddy et al. (2011) may need to access arbitrary parts of the data history in order to update the particles. Hence, although sequential inference is fast, the method is not technically *online*: tree complexity grows as $\log t$, and at every update each of the $\mathbf{x}^t = (\mathbf{x}_1, \dots, \mathbf{x}_t)$ locations are candidates for new splitting locations via *grow*. To enable online operation with constant memory requirements, this covariate pool (\mathbf{x}^t) must be reduced to a size w , constant in t . This can only be achieved via data discarding. Crucially, the analytic/parametric nature of DT leaves enables a large part of any discarded information to be retained in the form of informative leaf priors. In effect, this yields a *soft* implementation of data discarding, which we refer to as *datapoint retirement*. We show that retirement can preserve the posterior predictive properties of the tree even after data are discarded, and furthermore following subsequent *prune* and *stay* operations. The only situation where the loss of data hurts is when new data arrive and demand a more complex tree. In that case, any retired points would not be available as anchors for new partitions. Again, since tree operations are local in nature, only the small subtree nearby

$\eta(\mathbf{x}_t)$ is effected by this loss of DoFs, whereas the compliment $\mathcal{T}_t \setminus \eta(\mathbf{x}_t)$, i.e., most of the tree, is not affected.

3.1 Conjugate informative priors at the leaf level

Consider first a single leaf $\eta \in \mathcal{T}_t$ in which we have already retired some data. That is, suppose we have discarded $(\mathbf{x}_s, y_s)_{\{s\}}$ which was in η in $\mathcal{T}_{t'}$ at some time $t' \leq t$. The information in this data can be ‘remembered’ by taking the leaf-specific prior, $\pi(\theta_\eta)$, to be the posterior of θ_η given (only) the retired data. Suppressing the η subscript, we may take $\pi(\theta) =_{\text{df}} P(\theta \mid (\mathbf{x}_s, y_s)_{\{s\}}) \propto L(\theta; (\mathbf{x}_s, y_s)_{\{s\}}) \pi_0(\theta)$ where $\pi_0(\theta)$ is a baseline non-informative prior employed at all leaves. The *active data* in η , i.e., the points which have not been retired, enter into the likelihood in the usual way to form the leaf posterior.

It is fine to *define* retirement in this way, but more important to argue that such retired information can be updated losslessly, and in a computationally efficient way. Suppose we wish to retire one more datapoint, (\mathbf{x}_r, y_r) . Consider the following recursive updating equation:

$$\pi^{(\text{new})}(\theta) =_{\text{df}} P(\theta \mid (\mathbf{x}_s, y_s)_{\{s\}, r}) \propto L(\theta; \mathbf{x}_r, y_r) P(\theta \mid (\mathbf{x}_s, y_s)_{\{s\}}). \quad (3)$$

As shown below, the calculation in (3) is tractable whenever conjugate priors are employed.

Consider first the linear regression model, $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I})$, where $\mathbf{y} = (y_s)_{\{s\}}$ is the retired response data, and \mathbf{X} the retired *augmented* design matrix, i.e., whose rows are like $[1, \mathbf{x}'_s]'$, so that β_1 represents an intercept. With $\pi_0(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$, we obtain:

$$\pi(\beta, \sigma^2) =_{\text{df}} P(\beta, \sigma^2 \mid \mathbf{y}, \mathbf{X}) = \text{NIG}(\nu/2, s\nu/2, \beta, \mathcal{G}^{-1})$$

where NIG stands for Normal-Inverse-Gamma, and assuming the Gram matrix $\mathcal{G} = \mathbf{X}'\mathbf{X}$ is invertible and denoting $Xy = \mathbf{X}'\mathbf{y}$, $r = \mathbf{y}'\mathbf{y}$, we have $\nu = n - p$, $\beta = \mathcal{G}^{-1}Xy$, and $s^2 = \frac{1}{\nu}(r - \mathcal{R})$, where $\mathcal{R} = \beta' \mathcal{G} \beta$. Having discarded $(y_s, \mathbf{x}_s)_{\{s\}}$, we can still afford to keep in memory the values of the above statistics, as, crucially, their dimension does not grow with $|\{s\}|$. Updating the prior to incorporate an additional retiree (y_r, \mathbf{x}_r) is easy:

$$\mathcal{G}^{(\text{new})} = \mathcal{G} + X'_r X_r, \quad Xy^{(\text{new})} = Xy + \mathbf{X}'_r y_r, \quad s^{(\text{new})} = s + y'_r y_r, \quad \nu^{(\text{new})} = \nu + 1. \quad (4)$$

The constant leaf model may be obtained as a special case of the above, where $\mathbf{x}^* = 1$, $\mathcal{G} = \nu$ and $\beta = \mu$. For the multinomial model, the discarded response values y_s may be represented as indicator vectors \mathbf{z}_s , where $z_{js} = \mathbf{1}(y_s = j)$. The natural conjugate here is the *Dirichlet* $D(\mathbf{a})$. The hyperparameter vector \mathbf{a} may be interpreted as counts, and is updated in the obvious manner, namely $\mathbf{a}^{(\text{new})} = \mathbf{a} + \mathbf{z}_r$ where $z_{jm} = \mathbf{1}(y_r = j)$. A sensible baseline is $\mathbf{a}_0 = (1, 1, \dots, 1)$. See O’Hagan and Forster (2004) for more details.

Unfolding the updating equations (3) and (4) makes it apparent that retirement preserves the posterior distribution. Specifically, the posterior probability of parameters θ , given the active (non-retired) data still in η is

$$\pi(\theta \mid \mathbf{x}^\eta, y^\eta) \propto L(\theta; \mathbf{x}^\eta, y^\eta) \pi(\theta) \propto L(\theta; \mathbf{x}^\eta, y^\eta) L(\theta; (\mathbf{x}_s, y_s)_{\{s\}}) \pi_0(\theta) = L(\theta; \mathbf{x}^{\eta'}, y^{\eta'}) \pi_0(\theta),$$

where η' is η without having retired $(\mathbf{x}_s, y_s)_{\{s\}}$. Since the posteriors are unchanged, so are the posterior predictive distributions and the marginal likelihoods required for the SMC updates. Note that new data $(\mathbf{x}_{t+1}, y_{t+1})$ which do not update a particular node $\eta \in \mathcal{T}_t \rightarrow \mathcal{T}_{t+1}$ do not

change the properties of the posterior local to the region of the input space demarcated by η . It is as if the retired data were never discarded. Only where updates demand modifications of the tree local to η is the loss in DoF felt. We argue in Section 3.2 that this impact can be limited to operations which *grow* the tree locally. Cleverly choosing which points to retire can further mitigate the impact of discarding (see Section 4).

3.2 Managing informative priors at the tree level

Intuitively, DTs with retirement manage two types of information: a non-parametric memory comprising an active data pool of constant size $w \ll t$, which forms the leaf likelihoods; and a parametric memory consisting of possibly informative leaf priors. The algorithm we propose proceeds as follows. At time t , add the t^{th} datapoint to the active pool, and update the model by SMC exactly as explained in Section 2. Then, if t exceeds w , also select some datapoint, (\mathbf{x}_r, y_r) , and discard it from the active pool (see Section 4 for selection criteria), having first updated the associated leaf prior for $\eta(\mathbf{x}_r)^{(i)}$, for each particle $i = 1, \dots, N$, to ‘remember’ the information present in (\mathbf{x}_r, y_r) . This shifts information from the likelihood part of the posterior to the prior, exactly preserving the time- t posterior predictive distribution and marginal likelihood for every leaf in every tree.¹

The situation changes when the next data point $(\mathbf{x}_{t+1}, y_{t+1})$ arrives. Recall that the DT update chooses between *stay*, *prune*, or *grow* nearby each $\eta(\mathbf{x}_{t+1})^{(i)}$. Grow and prune moves are affected by the absence of the retired data from the active data pool. In particular, the tree cannot grow if there are no active data candidates to split upon. This informs our assessment of retiree selection criteria in Section 4, as it makes sense not to discard points in parts of the input space where we expect the tree to require further DoFs. Moreover, we recognise that the stochastic choice between the three DT moves depends both upon the likelihood, and retired (prior) information local to $\eta(\mathbf{x}_{t+1})^{(i)}$, so that the way that prior information propagates after a prune, or grow move, matters. The original DT model dictates how likelihood information (i.e., resulting from active data) propagates for each move. We must provide a commensurate propagation for the retired information to ensure that the resulting online trees stay close to their full data counterparts.

If a *stay* move is chosen stochastically, no further action is required: retiring data has no effect on the posterior. When nodes are grown or pruned, the retiring mechanism itself, which dictates how informative priors can salvage discarded likelihood information, suggests a method for splitting and combining that information. Following a *prune*, retired information from the pruned leaves, η and its sibling $S(\eta)$, must be pooled into the new leaf prior positioned at the parent $P(\eta)$. Conjugate updating suggests the following additive rule:

$$\mathcal{G}^{P(\eta)} = \mathcal{G}^\eta + \mathcal{G}^{S(\eta)}, \quad Xy^{P(\eta)} = Xy^\eta + Xy^{S(\eta)} \quad s^{P(\eta)} = s^\eta + s^{S(\eta)}, \quad \nu^{P(\eta)} = \nu^\eta + \nu^{S(\eta)}.$$

Note that this does not require access to the actual retired datapoints, and would result in the identical posterior even if the data had not been discarded.

A sensible *grow* move can be derived by reversing this logic. We suggest letting both novel child leaves $\ell(\eta)$ and $r(\eta)$ inherit the parent prior, but split its strength ν^η between

¹In fact, every data point under active management can [in a certain limited sense] be retired without information loss.

them at proportions equal to the active data proportions in each child. Let $\alpha = \frac{|\ell(\eta)|}{|\eta|}$. Then,

$$\begin{aligned} \nu_{\ell(\eta)} &= \alpha \nu_{\eta}, & \mathcal{G}^{\ell(\eta)} &= \alpha \mathcal{G}^{\eta}, & Xy^{\ell(\eta)} &= \alpha Xy^{\eta}, & s^{\ell(\eta)} &= \alpha s^{\eta}, \\ \nu_{r(\eta)} &= (1 - \alpha) \nu_{\eta}, & \mathcal{G}^{r(\eta)} &= (1 - \alpha) \mathcal{G}^{\eta}, & Xy^{r(\eta)} &= (1 - \alpha) Xy^{\eta}, & s^{r(\eta)} &= (1 - \alpha) s^{\eta}. \end{aligned}$$

In other words, the new child priors share the retired information of the parents with weight proportional to the number of active data points they manage relative to the parent. This preserves the total strength of retired information, preserves the balance between active data and parametric memory, and is *reversible*: subsequent *prune* operations will exactly undo the partitioned prior, combining it into the same prior sufficient statistics at the parent.

This brings to light a second cost to discarding data, the first being a loss of candidates for future partitioning. Nodes grown using priors built from retired points lack specific location information from the actual retired (\mathbf{x}_s, y_s) pairs. Therefore newly grown leaves must necessarily compromise between explaining the new data, e.g., $(\mathbf{x}_{t+1}, y_{t+1})$, with immediately local data active data to $\eta(\mathbf{x})_{t+1}$, and information from retired points with less localised influence. The weight of each component in the compromise is $|\eta|/(|\eta| + \nu_{\eta})$ and $\nu_{\eta}/(|\eta| + \nu_{\eta})$, respectively. Eventually as t grows, with $w \ll t$ staying constant, retired information naturally dominates, precluding new grows even when active partitioning candidates exist. This means that while the hierarchical way in which retired data filters through to inference (and prediction) at the leaves is sensible, it is doubly-important that data points in parts of the input space where the response is very complex should not be discarded.

4 Active discarding

It matters which data points are chosen for retirement, so it is desirable to retire datapoints that will be of “less” use to the model going forward. In the case of a drifting concepts, retiring *historically*, i.e., retiring the oldest datapoints, may be sensible. We address this in Section 5. Here we consider static concepts, or in other words i.i.d. data. We formulate the choice of which active data points to retire as an *active discarding* (AD) problem by borrowing (and reversing) techniques from the *active learning* (AL) literature. Regression and classification models separately, as they require different AD techniques. We shall argue that in both cases AD is, in fact, easier than AL since DTs enable thrifty analytic calculations not previously possible, which are easily updated within the SMC.

4.1 Active discarding for regression

Active learning (AL) procedures are sequential decision heuristics for choosing data to add to the design, usually with the aim of minimising prediction error. Two common AL heuristics are active learning MacKay (MacKay, 1992, ALM) and active learning Cohn (Cohn, 1996, ALC). They were popularised in the modern nonparametric regression literature (Seo et al., 2000) using GPs, and subsequently ported to DTs (Taddy et al., 2011). An ALM scheme selects new inputs \mathbf{x}^* with maximum variance for $y(\mathbf{x}^*)$, whereas ALC chooses \mathbf{x}^* to maximise the expected reduction in predictive variance averaged over the input space. Both approximate maximum expected information designs in certain cases. ALC is computationally more demanding than ALM, requiring an integral over a set of reference locations that

can be expensive to approximate numerically for most models. But it leads to better exploration when used with nonstationary models like DTs because it concentrates sampled points near to where the response surface is changing most rapidly (Taddy et al., 2011). ALM has the disadvantage that it does not cope well with heteroskedastic data (i.e., input-dependent noise). It can end up favouring regions of high noise rather than high model uncertainty. Both are sensitive to the choice of (and density of) a search grid over which the variance statistics are evaluated.

Our first simplification when porting AL to AD is to recognise that no grids are needed. We focus on the ALC statistic here because it is generally preferred, but also to illustrate how the integrals required are actually very tractable with DTs, which is not true in general. The AD program is to evaluate the ALC statistic at each active data location, and choose the smallest one for discarding. AL, by contrast, prefers large ALC statistics to augment the design. We focus on the linear leaf model, as the constant model may be derived as a special case. For an active data location \mathbf{x} and (any) reference location \mathbf{z} , the reduction in variance at \mathbf{z} given that \mathbf{x} is in the design, and a tree \mathcal{T} is given by (see Taddy et al. (2011)):

$$\Delta\sigma_{\mathbf{x}}^2(\mathbf{z}|\mathcal{T}) = \Delta\sigma_{\mathbf{x}}^2(\mathbf{z}|\eta) \equiv \sigma^2(\mathbf{z}|\eta) - \sigma_{\mathbf{x}}^2(\mathbf{z}|\eta) = \frac{s_{\eta}^2 - \mathcal{R}_{\eta}}{|\eta| - m - 3} \times \frac{\left(\frac{1}{|\eta|} + \mathbf{z}'\mathcal{G}_{\eta}^{-1}\mathbf{x}\right)^2}{1 + \frac{1}{|\eta|} + \mathbf{x}'\mathcal{G}_{\eta}^{-1}\mathbf{x}},$$

when *both* \mathbf{x} and \mathbf{z} are in $\eta \in \mathcal{L}_{\mathcal{T}}$, and zero otherwise. This expression is valid whether learning or discarding, however AL requires evaluating $\Delta\sigma^2(x)$ over a dense candidate grid of x 's. AD need only consider the current active data locations, which can represent a dramatic savings in computational cost.

Integrating over \mathbf{z} gives:

$$\Delta\sigma^2(\mathbf{x}) = \int_{\mathbb{R}^d} \Delta\sigma_{\mathbf{x}}^2(\mathbf{z}) d\mathbf{z} = \frac{s_{\eta}^2 - \mathcal{R}_{\eta}}{(|\eta| - m - 3)(1 + \frac{1}{|\eta|} + \mathbf{x}'\mathcal{G}_{\eta}^{-1}\mathbf{x})} \times \int_{\eta} \left(\frac{1}{|\eta|} + \mathbf{z}'\mathcal{G}_{\eta}^{-1}\mathbf{x}\right)^2 d\mathbf{y}.$$

The integral that remains, over the rectangular region η , is tedious to write out but has a trivial $O(m^2)$ implementation. Let the m -rectangle η be described by $\{(a_i, b_i)\}^m$. Then,

$$\begin{aligned} \int_{a_1}^{b_1} \cdots \int_{a_m}^{b_m} \left(c + \sum_{i=1}^m \tilde{z}_i x_i\right)^2 dz_1 \cdots dz_m &= A_{\eta} c^2 + c \sum_i \left(\prod_{k \neq i} (b_k - a_k)\right) x_i (b_i^2 - a_i^2) \\ &+ \sum_i \left(\prod_{k \neq i} (b_k - a_k)\right) \frac{x_i^2}{3} (b_i^3 - a_i^3) + \sum_i \sum_{j < i} \left(\prod_{k \neq i, j} (b_k - a_k)\right) \frac{x_i x_j}{2} (b_i^2 - a_i^2)(b_j^2 - a_j^2), \end{aligned}$$

where $\tilde{\mathbf{z}} = \mathbf{z}'\mathcal{G}_{\eta}^{-1}$, and $c = 1/|\eta|$. A general-purpose numerical version via sums using R reference locations \mathbf{z} —previously the state of the art (Taddy et al., 2011)—requires $O(Rm)$ computation with R growing exponentially in m for reasonable accuracy. Observe that the rectangular leaf regions generated by the trees is key. In the case of other partition models (like Voronoi tessellation models), this analytical integration would not be possible.

In repeated applications of ALC for AD, we observe that the active points that remain tend shuffle themselves so that they cluster near the high posterior partitioning boundaries, which makes sense because these are the locations where the predictive surface is changing the

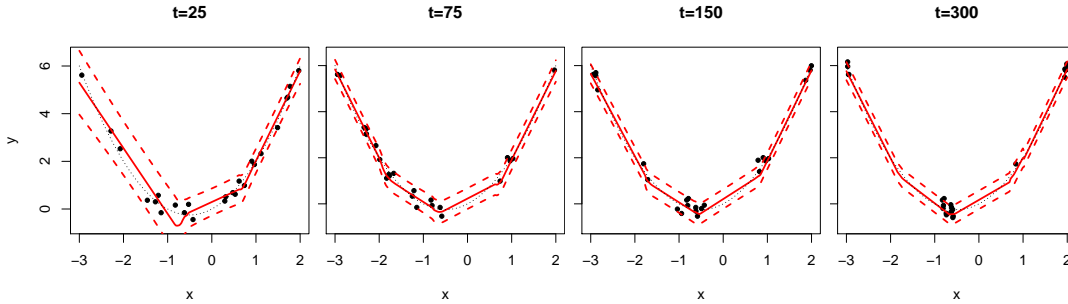


Figure 1: Snapshots of active data (25 points) and predictive surfaces spanning 275 retirement/updating rounds; $Y(x) = x + x^2 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 1)$.

fastest. The number of such locations depends on the number of active data points allowed, w . As an illustration, consider the simple example where the response is a parabolic function, which must be learned sequentially via x -data sampled uniformly in $(-3, 2)$, with $w = 25$. The initial 25, before any retiring, are shown in the first panel. Each updating round then proceeds with one retirement followed by one new pair, and subsequent SMC update. Since the implementation requires at least five points in each leaf, seeing four regimes emerge is perhaps not surprising. By $t = 150$, the third pane, the ability to learn about the mean with just 25 degrees of freedom is saturated, but it is possible to improve on the variance (shown as errorbars), which are indeed smaller in the final $t = 300$ pane. Eventually, the points will cluster at the ends because that is where the response is changing most rapidly, and indeed the derivative is highest there (in absolute value).

4.2 Active discarding for classification

For classification, predictive entropy is an obvious AL heuristic. Given a predictive surface comprised of probabilities $p_\ell(\mathbf{x})$ for each class ℓ , from DTs or otherwise, the predictive entropy at \mathbf{x} is $-\sum_\ell p_\ell(\mathbf{x}) \log p_\ell(\mathbf{x})$. Entropy can be an optimal method for measuring predictive uncertainty, but that does not mean it is good for AL. Many authors (e.g., Joshi et al., 2009) have observed that it can be too greedy: entropy can be very high near the best explored class boundaries. Several, largely unsatisfactory, remedies have been suggested in the literature. Fortunately, no remedy is required for the AD analog, which focuses on the lowest entropy active data, a finite set. The discarded points will tend to be far into the class interior, where they can be safely subsumed into the prior. Their spacing and shifting of the active pool is quite similar to discarding by ALC for regression, and so we do not illustrate it here.

4.3 Fast local updates of active discarding statistics with trees

The divide and conquer nature of trees—whose posterior distribution is approximated by thrifty, local, particle updates—allows AD statistics to be updated cheaply too. If each leaf node stores its own AD statistics, it suffices to update only the ones in leaf nodes which have been modified, as described below. Any recalculated statistics can then be subsumed into

a global, particle averaged, version. Note that no updates to the AD statistics are needed when a point is retired since the predictive distributions are unchanged.

When a new datapoint (\mathbf{x}, y) arrives, the posterior undergoes two types of changes: resample then propagate. In the resample step the discrete particle distribution changes, although the trees therein do not change. Therefore, each discarded particle must have its AD statistics (stored at the leaves) subtracted from the full particle tally. Then each correspondingly duplicated particle can have its AD statistics added in. No new integrals (for ALC) or entropy calculations (for classification) are needed. In the propagate step, each particle undergoes a change local to $\eta(\mathbf{x})^{(i)} \in \mathcal{T}_t^{(i)}$. This requires first calculating the AD statistic for the new (\mathbf{x}, y) for each $\eta^{(i)}(\mathbf{x})$, before the dynamic update occurs, and then swapping it into the particle average. New integrations, etc., need evaluating here. Then, each non-*stay* dynamic update triggers swap of the old AD statistics in $\eta^{(i)}(\mathbf{x})$ for freshly re-calculated ones from the leaf node(s) in $\mathcal{T}_{t+1}^{(i)}$. The total computational cost is in $O(m^2 N)$ for incorporating (\mathbf{x}, y) into N particles, plus $O(m^2 \sum_{i=1}^N |\eta^{(i)}(x)|)$ to update the leaves.²

4.4 Empirical results

Here we explore the benefit of AD over simpler heuristics, like random discarding and sub-setted data estimators, by making predictive comparisons on benchmark regression and classification data. To focus the discussion on our key objective for this section, we employ moderate data sample sizes in order to allow a comparison to full-data versions of DTs, and assess the impact of data discarding on performance. In particular, we do not repeat here a comparison of full-data DTs to competitors, which may be found in Taddy et al. (2011), but emphasise that discarding enables DTs to operate on (arbitrarily long) data streams, where the original DTs, as well as their main GP-based competitors, will eventually become intractable. This is better illustrated by the use of massive and streaming classification datasets in Section 5.

Simple synthetic regression data

We first consider data originally used to illustrate multivariate adaptive regression splines (MARS) (Friedman, 1991), and then to demonstrate the competitiveness of DTs relative to modern (batch) nonparametric models (Taddy et al., 2011). The response is $10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$ plus $\mathcal{N}(0, 1)$ additive error. Inputs \mathbf{x} are random in $[0, 1]^5$. We considered four estimators: one based on 200 pairs (ORIG), one based on 1800 more for 2000 total (FULL), and two online versions using either random (ORAND) or ALC (OALC) retiring to keep the total active data set limited to $w = 200$. ORIG is intended as a lower benchmark, representing a naïve fixed-budget method; FULL is at the upper end. The full experiment was comprised of 100 repeats in a MC fashion, each with new random training sets, and random testing sets of size 1000. $N = 1000$ particles and a linear leaf model were used throughout. Similar results were obtained for the constant model.

Figure 2 reveals that random retiring is better than subsetting, but retiring by ALC is even better, and can be nearly as good as the full-data estimator. In fact, OALC was the *best*

²One might imagine a thriftier, but harder to implement version, which waits until the end to calculate the AD statistic for the new point (x, y) . But it would have the same computational order.

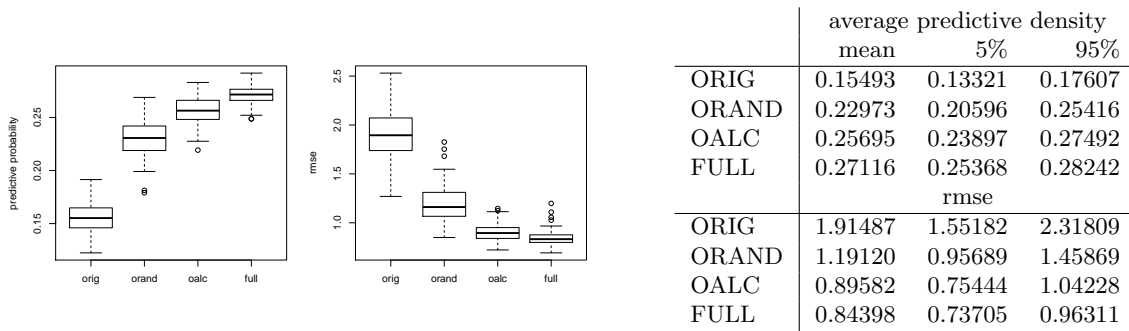


Figure 2: Friedman data comparisons by average posterior predictive density (higher is better) and RMSE (lower is better).

predictor 16% and 28% of the time by average predictive density and RMSE, respectively. The average time used by each estimator was approximately 1, 33, 45, and 67 seconds, respectively. So random retiring on this modestly-sized problem is 2-times faster than using the full data. ALC costs about 18% extra, time-wise, but leads to about a 35% reduction in RMSE relative to the full estimator. We note that in much larger problems the gap between the online and full estimators can widen considerably. The time-demands of the full estimator grow roughly as $t \log t$, whereas the online versions stay constant.

Spam classification data

Now consider the Spambase data set, from the UCI Machine Learning Repository (Asuncion and Newman, 2007). The data contains binary classifications of 4601 emails based on 57 attributes (predictors). We report on a similar experiment to the Friedman/regression example, above, except with classification leaves and 5-fold CV to create training and testing sets. This was repeated twenty times, randomly, giving 100 sets total. Again, four estimators were used: one based on 1/10 of the training fold (ORIG), one based on the full fold (FULL), and two online versions trained on the same stream(s) using either random (ORAND) or entropy (OENT) retiring to keep the total active data set limited to 1/10 of the full set.

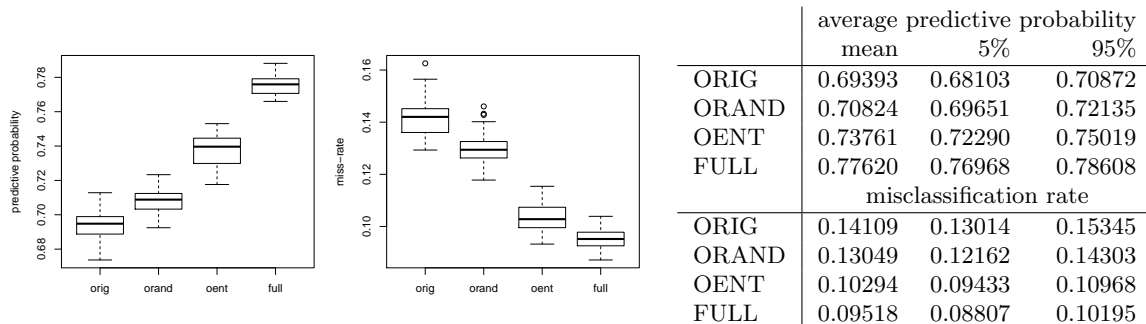


Figure 3: Spam data comparisons by average posterior predictive probability (higher is better) and misclassification rate (lower is better) on the testing set(s).

Figure 3 tells a similar story to the Friedman experiment: random discarding is better than subsetting, but discarding by entropy is even better, and can be nearly as good as

the full-data estimator. Entropy retiring resulted the best predictor 7% of the time by misclassification rate, but never by posterior predictive probability.

5 Temporal adaptivity using forgetting factors

The accumulation of historical information at the leaf priors introduced by data retirement may eventually overpower the likelihood of active datapoints. This is natural in an i.i.d. setting, but may cause performance deterioration in streaming contexts where the data generating mechanism may evolve or change suddenly. To promote responsiveness, we may *exponentially downweight* the retired data history s when retiring an additional point y_m : $\pi_\lambda^{(\text{new})}(\theta) \propto L(\theta \mid y_m)L^\lambda(\theta; (y_s, \mathbf{x}_s)_{\{s\}})\pi_0(\theta)$. Observe that for $\lambda = 1$, conjugate Bayesian updating is recovered, and for $\lambda = 0$, the retired history is disregarded altogether, effectively resetting the prior. For $\lambda \in (0, 1)$, two effects are introduced. First, the overall ‘strength’ of the prior relative to the likelihood is diminished. Second, as the prior is sequentially updated, it will place disproportionately more weight on recently retired datapoints as opposed to older retired data. For the leaf models entertained in this paper, a recursive application of this principle, with $\lambda \in (0, 1)$, modifies only slightly the conjugate updates of Section 3, as follows. For the linear and constant models, we have $(\mathbf{A}^{(\text{new})})^{-1} = \lambda \mathbf{A}^{-1} + X'_m X_m$, $\mathbf{R}^{(\text{new})} = \lambda \mathbf{R} + \mathbf{X}'_m \mathbf{y}_m$, $s^{(\text{new})} = \lambda s + y'_m y_m$, and $\nu^{(\text{new})} = \lambda \nu + 1$, whereas for the multinomial, we get $\mathbf{a}^{(\text{new})} = \lambda \mathbf{a} + \mathbf{z}_m$. For $\lambda < 1$, κ and ν will be bounded above by their limiting value $\frac{1}{1-\lambda}$, irrespective of the total number of retired datapoints.

In Ibrahim et al. (2003), this family of priors is shown to satisfy desirable information-theoretic optimality properties. Exponential downweighting as a means of enabling temporal adaptivity also has a long tradition in non-stationary signal processing (Haykin, 1996), as well as streaming classification (Anagnostopoulos et al., 2009), where λ is often referred to as a *forgetting factor*.

In historical discarding, it is perhaps obvious that some degree of forgetting will be useful in drifting contexts, as the contribution of past data becomes decreasingly useful with time. The relationship between forgetting and other types of active discarding is however more complex. In principle, any successful active discarding scheme will lead to priors being populated by less relevant datapoints, so that the model can benefit from forgetting in favour of putting more weight on highly relevant, active data. Unfortunately, in the presence of drift, we cannot guarantee such reasonable behaviour from active discarding heuristics of the form proposed here. As these latter are reliant on an i.i.d. assumption, they can often mistake obsolete datapoints that are poorly explained by the model for ‘highly informative, surprising’ datapoints that had better be retained, so that it becomes less clear *a priori* whether the active data pool or the prior should be ‘trusted’ more, and the utility of forgetting becomes questionable. More sophisticated active learning heuristics are required to resolve this problem, which lie beyond the scope of this paper. We will thus only explore the interaction of forgetting with historical discarding henceforth.

5.1 Synthetic drifting regression data

We now revisit the Friedman dataset from 4.4, and introduce smooth drift by replacing the non-linear term $10 \sin(\pi x_1 x_2)$ with a time-varying version, $10a_t \sin(\pi x_1 x_2)$. The coefficient a_t

is allowed to vary smoothly between -1 and 3 over time as $a_t = 2 \sin(2\pi kt/1000) + 1$, so that k controls the speed of the drift: $k = 1$ producing one full cycle every 1000 timesteps. Note that as a_t increases in magnitude, the non-linearity of the regression surface will accentuate, as the first term is responsible for much of its complexity. The simulation measures 1-step-ahead performance of the DT as follows: at each timestep t , it first generates 5 datapoints from the current model; these are used as test datapoints to measure the predictive probability and RMSE of the dynamic tree (trained using data up to time $t - 1$); and, finally, the DT is updated on the basis of the new data.

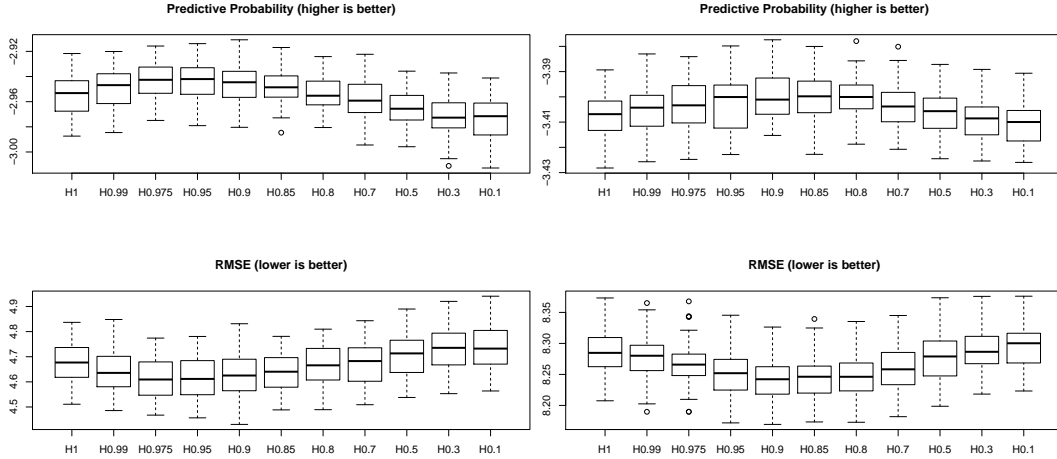


Figure 4: Slowly (left) and rapidly (right) drifting Friedman data comparisons by average posterior predictive density (top - higher is better) and RMSE (bottom - lower is better), for various degrees of forgetting.

In the Figure 4 plots, we plot the RMSE and predictive probability observed over 100 MC iterations for a sequence of λ values between $\lambda = 0$ (discarding with retiring) and $\lambda = 1$ (retirement via Bayesian conjugate updating). Reassuringly, a U-shaped curve appears, indicating a trade-off between throwing away too much information at one extreme ($\lambda = 0$), and retaining obsolete information at the other ($\lambda = 1$). For rapidly changing data distributions ($k = 1$), a value of $\lambda = 0.8$ seems to perform best. Repeating the experiment for slower-changing data distributions ($k = 0.1$) produces performance that peaks at $\lambda = 0.97$ instead, confirming our intuition.

In Figure 5, we investigate the effect that discarding and forgetting have on model complexity, as measured by average tree height over time. To do so, we again generate drifting Friedman data, this time with $a_t = 10$ between $t = 10000$ and $t = 20000$ (denoted by vertical lines in Figure 5), and 0 otherwise so that model complexity rises sharply and then drops again. We deploy a DT without discarding (i.e., sequentially incorporating the full dataset), a DT with a fixed budget of 100 active datapoints and no forgetting, and one with the same budget and mild forgetting ($\lambda = 0.9$). First observe that capping the active data pool size significantly penalises model complexity on the whole. Also note that all three methods react to the rise in complexity at $t = 10^4$ by favouring deeper trees. However, once the data complexity drops again at $t = 2 \times 10^4$, both the full model and the online model without forgetting ($\lambda = 1$) retain their average tree depth, failing to return to earlier levels. By

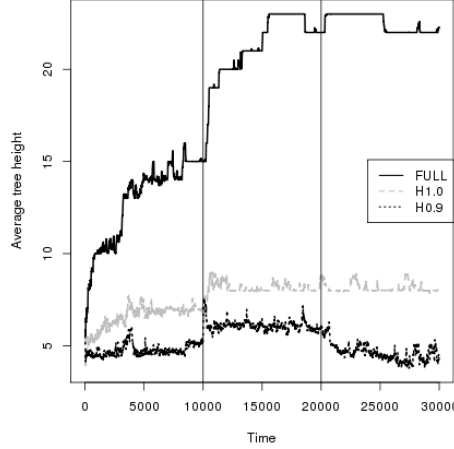


Figure 5: Average tree height over time for the full model, and historical discarding with $\lambda = 0.9$ or $\lambda = 1$. The true regression surface complexity rises in $t \in [10^4, 2 \times 10^4]$.

contrast, $\lambda = 0.9$ allows the model to adapt to the change, as the priors are more easily outweighed by the impact of novel information.

5.2 Synthetic drifting classification data

We now turn to streaming classification. We henceforth adopt the standard one-step-ahead performance assessment paradigm, wherein the algorithm at time t first predicts the class label of the unlabelled $(t + 1)$ th datapoint, and is then allowed to use both the datapoint itself and its label to update its parameters.

We first consider a classification problem where the optimal decision boundary is always non-linear, but drifts in time in such a way so that older data become increasingly misleading for future predictions. This effect can be synthesised by rotating a ‘fuzzy’ XOR problem, displayed in the left plot of Figure 6. The XOR forces a non-linear decision boundary, whereas the rotation implies that recent data should have priority over older data. This example, which we refer to as MOVINGTARGET, is an extreme one since in general drift could also manifest itself in ways that render past information useless, but not outright misleading. Even in such cases, data discarding and forgetting may be useful to ‘free up’ degrees of freedom, but the effect is unlikely to be as dramatic and would therefore be harder to measure.

In the right plot of Figure 6 and the rightmost column of Table 1, we illustrate the effect that introducing a forgetting factor has on the performance of a DT with historical discarding against MOVINGTARGET. We compare against two state-of-the-art methods, Quadratic Discriminant Analysis with Adaptive Forgetting (QDA-AF) (Anagnostopoulos et al., 2009), and Online Linear Discriminant Analysis with Adaptive Learning Rate (OLDA-ALR) (Kuncheva and Plumptre, 2008), both designed with streaming classification contexts in mind. In Table 1, performance is measured in three ways: correct classification rate, Area Under the Curve, the H -measure, a newly preferred alternative to the AUC (Hand, 2009).

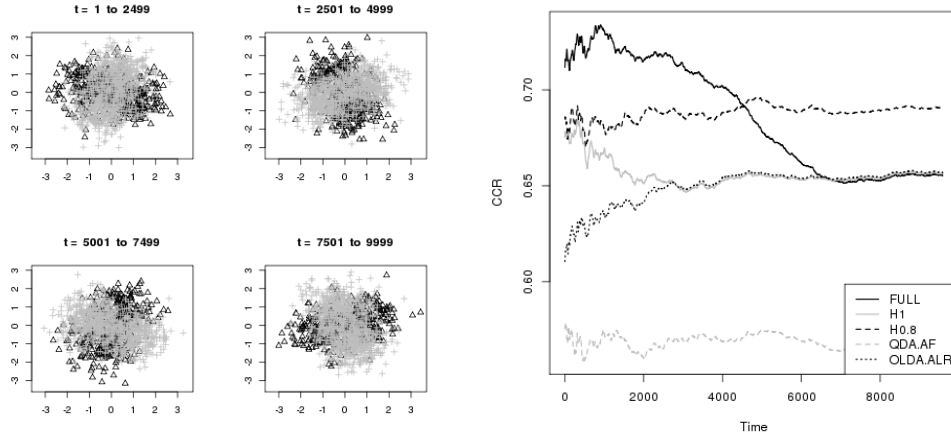


Figure 6: Left: four snapshots of the drifting synthetic classification data underlying this simulation study. Right: average classification performance (Correct Classification Rate) over time.

Table 1: Performance in terms of Area Under the Curve (AUC), H-measure (H) and Correct Classification Rate (CCR) for each of three datasets (two real and one simulated), and three instantiations of dynamic trees as well as two competitive methods.

	ELEC2			FAUD			MOVINGTARGET		
	AUC	H	CCR	AUC	H	CCR	AUC	H	CCR
OFFLINE ($n = 10^4$)	0.771	0.267	0.702	0.588	0.048	0.941	0.562	0.044	0.655
ONLINE ($\lambda = 1$)	0.761	0.274	0.724	0.724	0.155	0.971	0.528	0.011	0.656
ONLINE ($\lambda = 0.8$)	0.880	0.480	0.808	0.930	0.622	0.982	0.668	0.111	0.609
QDA.AF	0.924	0.643	0.873	0.973	0.920	0.983	0.504	0.001	0.560
OLDA.AL	0.763	0.239	0.692	0.832	0.414	0.974	0.507	0.001	0.656

In all three respects, data discarding hardly improves performance when $\lambda = 1$, whereas for $\lambda = 0.9$ significant improvement is possible. For an explanation, consider the way in which classification performance evolves over time, shown on the rightmost plot of Figure 6: the detrimental effect of obsolete, misleading data becomes visually obvious for the full-data model, as well as the online model with $\lambda = 1$. Interestingly, these latter two methods, although otherwise distinct, share a similar performance bottleneck with OLDA-ALR that DTs with forgetting to decidedly overcome.

Now consider two real datasets, ELEC2, and FRAUD, which are known to exhibit concept drift (Anagnostopoulos et al., 2009). The former holds information for the Australian New South Wales Electricity Market and was introduced in Baena-Garcia et al. (2006), comprising 27552 instances, each referring to a period of 30 minutes. The class label identifies the price change related to a moving average of the last 24 hours, and the four covariates capture aspects of electricity demand and supply. The latter dataset, FRAUD, is of length prohibitive to many existing methods ($n = 100,000$), and contains information about credit card transactions, and their respective status as legitimate or fraudulent, determined by experts (see Anagnostopoulos et al. (2009) for more details). Results in terms of CCR over

time are presented in Figure 7. The full suite of numerical results is provided in Table 1.

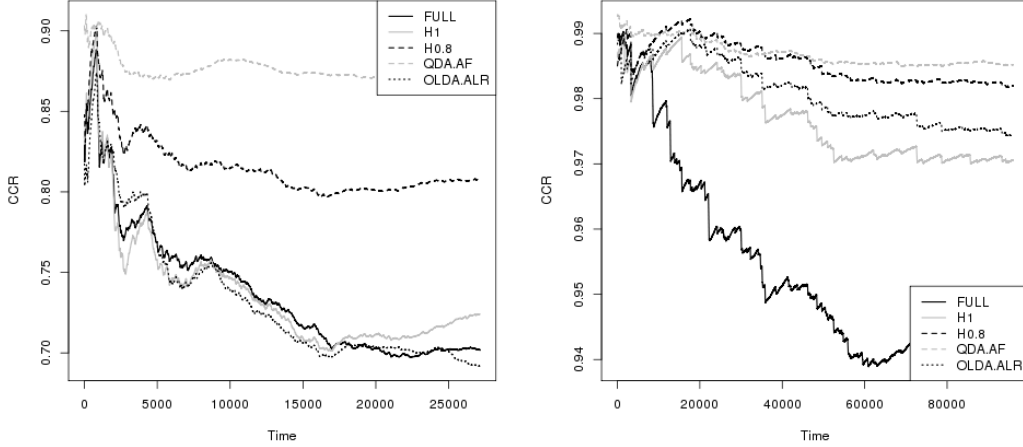


Figure 7: Left: Electricity Market data. Right: Fraud data. Average classification performance (Correct Classification Rate) over time.

QDA-AF, which was vastly inferior in the synthetic MOVINGTARGET example, dominates in these examples (particularly for FRAUD). Comparisons between the remaining methods paint an interesting picture. In both datasets, FULL is among the lowest performers, suggesting that data discarding is essential to maintaining a representative fit against drifting data distributions. Among DTs, $\lambda = 1$ performs poorly, as informative priors accumulate irrelevant or misleading information. In contrast, $\lambda = 0.8$ is much better, outperforming OLDA-ALR. Although the ranking of various classifiers will generally differ by application, these experiments give a strong signal that the use of forgetting factors can turn DTs into a promising, flexible tool for streaming classification.

6 Conclusion

In this work, we strive to fully utilise the potential of Bayesian machinery in the context of streaming non-parametrics. We propose data retirement via conjugate Bayesian updating in the context of SMC inference for a dynamic tree model, preserving non-parametric flexibility while enabling constant memory online operation. Second, the availability of tractable predictive distributions allows us to devise computationally efficient active retirement heuristics, hence maintaining a fixed budget of highly informative datapoints. Both features minimise information loss incurred by single-pass processing. Finally, we deploy informative power priors to enable temporal adaptivity. This results in a novel, powerful algorithmic scheme for non-parametric regression and classification tailored to the massive and streaming data contexts. As future work, we intend to pursue techniques for automatic tuning of forgetting factors in streaming contexts, and their interplay with active retirement heuristics.

7 Acknowledgements

The first author was supported by a Cambridge Statistics Initiative Research Fellowship at the Statistical Laboratory, University of Cambridge, for a large part of this work.

References

- Anagnostopoulos, C., N. Adams, N. Pavlidis, D. Tasoulis, and D. Hand (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning* 77, 103–123.
- Asuncion, A. and D. Newman (2007). UCI machine learning repository.
- Baena-Garcia, M., J. del Campo-Avila, R. Fidalgo, A. Bifet, R. Gavalda, and R. Morales-Bueno (2006). Early drift detection method. In *ECML PKDD 2006 Workshop on Knowledge Discovery from Data Streams*, pp. 77–86.
- Carvalho, C. M., M. Johannes, H. F. Lopes, and N. G. Polson (2010). Particle learning and smoothing. *Statistical Science* 25, 88–106.
- Chipman, H., E. George, and R. McCulloch (1998). Bayesian CART model search (with discussion). *Journal of the American Statistical Association* 93, 935–960.
- Chipman, H., E. George, and R. McCulloch (2002). Bayesian treed models. *Machine Learning* 48, 303–324.
- Cohn, D. A. (1996). Neural network exploration using optimal experimental design. In *Advances in Neural Information Processing Systems*, Volume 6(9), pp. 679–686. Morgan Kaufmann Publishers.
- Friedman, J. H. (1991, March). Multivariate adaptive regression splines. *Annals of Statistics* 19, No. 1, 1–67.
- Gramacy, R. B. and M. A. Taddy (2011). *dynaTree: Dynamic trees for learning and design*. Booth School of Business, University of Chicago. R package version 2.0.
- Hand, D. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning* 77, 103–123.
- Haykin, S. (1996). *Adaptive Filter Theory*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA.
- Ibrahim, J., M. Chen, and D. Sinha (2003). On optimality properties of the power prior. *Journal of the American Statistical Association* 98(461), 204–213.
- Joshi, A., F. Porikli, and N. Papanikolopoulos (2009). Multi-class active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. to appear.
- Kuncheva, L. and C. Plumpton (2008). Adaptive learning rate for online linear discriminant classifiers. *SSPR and SPR 2008, Lecture Notes in Computer Science (LNCS)* 5342, 510–519.

- MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation* 4(4), 589–603.
- O’Hagan, A. and J. Forster (2004). *Kendall’s Advanced Theory of Statistics, Volume 2B, Bayesian Inference*. Arnold Publishers.
- Seo, S., M. Wallat, T. Graepel, and K. Obermayer (2000, July). Gaussian process regression: Active data selection and test point rejection. In *Proceedings of the International Joint Conference on Neural Networks*, Volume III, pp. 241–246. IEEE.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Taddy, M., R. Gramacy, and N. Polson (2011). Dynamic trees for learning and design. *Journal of the American Statistical Association* 106(493), 109–123.